

The new CMS DAQ system for LHC operation after 2014 (DAQ2)

This content has been downloaded from IOPscience. Please scroll down to see the full text.

2014 J. Phys.: Conf. Ser. 513 012014

(<http://iopscience.iop.org/1742-6596/513/1/012014>)

View [the table of contents for this issue](#), or go to the [journal homepage](#) for more

Download details:

IP Address: 193.198.162.14

This content was downloaded on 16/08/2016 at 14:03

Please note that [terms and conditions apply](#).

You may also be interested in:

[Upgrade of the CMS Event Builder](#)

G Bauer, U Behrens, M Bowen et al.

[Studies on load metric and communication for a load balancing algorithm in a distributed data acquisition system](#)

M Simon, H Sakulin and S Kozielski

[10 Gbps TCP/IP streams from the FPGA for the CMS DAQ eventbuilder network](#)

G Bauer, T Bawej, U Behrens et al.

[10 Gbps TCP/IP streams from the FPGA for High Energy Physics](#)

Gerry Bauer, Tomasz Bawej, Ulf Behrens et al.

[Operational experience with the CMS Data Acquisition System](#)

G Bauer, U Behrens, M Bowen et al.

[Dynamic configuration of the CMS Data Acquisition cluster](#)

G Bauer, U Behrens, K Biery et al.

[Prototype of a File-Based High-Level Trigger in CMS](#)

G Bauer, T Bawej, U Behrens et al.

The new CMS DAQ system for LHC operation after 2014 (DAQ2)

Gerry Bauer⁶, Tomasz Bawe², Ulf Behrens¹, James Branson⁴,
Olivier Chaze², Sergio Cittolin⁴, Jose Antonio Coarasa²,
Georgiana-Lavinia Darlea⁶, Christian Deldicque², Marc Dobson²,
Aymeric Dupont², Samim Erhan³, Dominique Gigi², Frank Glege²,
Guillermo Gomez-Ceballos⁶, Robert Gomez-Reino², Christian Hartl²,
Jeroen Hegeman², Andre Holzner⁴, Lorenzo Masetti², Frans
Meijers², Emilio Meschi², Remigius K Mommsen⁵, Srecko Morovic^{2,7},
Carlos Nunez-Barranco-Fernandez², Vivian O'Dell⁵, Luciano Orsini²,
Wojciech Ozga², Christoph Paus⁶, Andrea Petrucci², Marco Pieri⁴,
Attila Racz², Olivier Raginel⁶, Hannes Sakulin², Matteo Sani⁴,
Christoph Schwick², Andrei Cristian Spataru², Benjamin Stieger²,
Konstanty Sumorok⁶, Jan Veverka⁶, Christopher Colin Wakefield² and
Petr Zejdl²

¹ DESY, Hamburg, Germany

² CERN, Geneva, Switzerland

³ University of California, Los Angeles, Los Angeles, California, USA

⁴ University of California, San Diego, San Diego, California, USA

⁵ FNAL, Chicago, Illinois, USA

⁶ Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

⁷ Now at Institute Rudjer Boskovic, Zagreb, Croatia

E-mail: Andre.Georg.Holzner@cern.ch

Abstract. The Data Acquisition system of the Compact Muon Solenoid experiment at CERN assembles events at a rate of 100 kHz, transporting event data at an aggregate throughput of 100 GByte/s. We are presenting the design of the 2nd generation DAQ system, including studies of the event builder based on advanced networking technologies such as 10 and 40 Gbit/s Ethernet and 56 Gbit/s FDR Infiniband and exploitation of multicore CPU architectures. By the time the LHC restarts after the 2013/14 shutdown, the current compute nodes, networking, and storage infrastructure will have reached the end of their lifetime. In order to handle higher LHC luminosities and event pileup, a number of sub-detectors will be upgraded, increase the number of readout channels and replace the off-detector readout electronics with a μ TCA implementation. The second generation DAQ system, foreseen for 2014, will need to accommodate the readout of both existing and new off-detector electronics and provide an increased throughput capacity. Advances in storage technology could make it feasible to write the output of the event builder to (RAM or SSD) disks and implement the HLT processing entirely file based.



1. Introduction

The central data acquisition system (DAQ) of the Compact Muon Solenoid (CMS) at the Large Hadron Collider (LHC) at CERN has delivered an excellent performance during LHC run 1 [1]. However, the demands on the DAQ system for LHC run 2 after the Long Shutdown 1 (LS1) will be such that the current system will not meet them (mainly but not exclusively due to the increased instantaneous luminosity, especially if LHC has to continue to run at 50ns bunch spacing). In addition, most of the DAQ equipment has reached the end of the 5 year replacement cycle. CMS will therefore build a new DAQ system which will also accommodate new μ TCA [2] based Front End Drivers (FEDs) where the Slink64 [3] connection to the central DAQ has been replaced by an optical link with more bandwidth.

More details on the requirements for the new DAQ system are given in section 2 while section 3 describes the proposed layout. The new interface to the FEDs is described in [4]. Section 4 contains a summary of the next part in the data path, the event builder core. A few notes on the necessary performance tuning for such a system can be found in section 5 which also discusses the advantages of Infiniband used in the event builder core. The new interface to the High Level Trigger (HLT) system is described in [5]. Results obtained with a small scale demonstrator system are described in section 6. Finally, section 7 concludes the article and section 8 gives an outlook.

2. Requirements

Table 1 shows the comparison of the requirements for the DAQ1 and DAQ2 systems to highlight the differences and commonalities between them.

The overall readout capacity needs to be increased as we need to be able to handle larger event sizes compared to run 1 due to increased instantaneous luminosity and additional detector readout channels. Also, a new interface (Slink-Express) to new Front End Drivers must be supported. A further new requirement is the decoupling of the event builder software from the high level trigger software. This is discussed in more detail in [5].

Table 1. Comparison of requirements for DAQ1 and DAQ2

requirement	DAQ1	DAQ2
readout rate	100 kHz	100 kHz
Front End Drivers	640: 1 - 2 kByte	640: 1 - 4 kByte 50: 2 - 8 kByte
total readout bandwidth	100 GByte/s	200 GByte/s
interface to FEDs	Slink64	Slink64/Slink-Express
coupling of event builder software to HLT software	no requirement	decoupled
lossless event building	required	required
HLT capacity	extendable	extendable
high availability / fault tolerance	required	required
cloud facility for offline processing	originally not required	required

3. DAQ2 layout

The general layout of the DAQ2 system (see figure 1) is similar to the one of DAQ1. The subdetector data enter the central DAQ system via the Slink64 or Slink-Express for new FEDs

point-to-point connections to the Front End Readout Optical Links (FEROLs) [4] installed in the service cavern underground. The FEROLs convert the data stream to TCP/IP and therefore act as interface between the custom CMS-specific hardware and commercially available equipment. Through optical 10 Gbit/s Ethernet links the data are brought to the surface via a layer of 10/40 Gbit/s Ethernet switches. Readout Unit PCs (RUs) assemble the fragments from several FEROLs into superfragments which are sent via Infiniband (see section 5.2) to Builder Unit PCs (BUs) for assembly of the entire event (see section 4). Several Filter Unit PCs (FUs) are assigned to each Builder Unit. The Filter Units analyze the fully assembled events made available by the corresponding Builder Unit and take the high level trigger decision. Accepted events are written to files on a cluster file system which are then transferred to the Tier0 computing facility for offline processing.

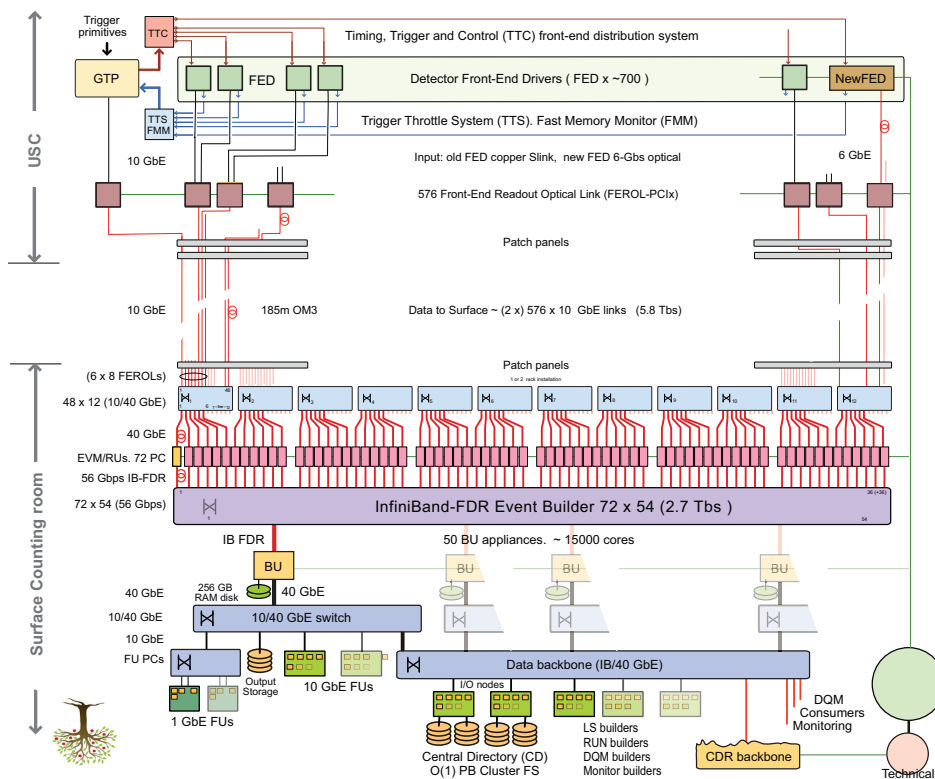


Figure 1. DAQ2 proposed layout. The data flows from the top to the bottom, the upper part (Underground Service Cavern, USC) will be installed underground, while the lower part will be in the surface computing room.

4. Event builder core

CMS has used a two layered event builder in LHC run 1: the first layer builds ‘superfragments’ from eight to sixteen event fragments while the assembly of the full event (from the superfragments) is done only in the second layer. This was mainly imposed by the lack of availability of cost effective large ($\sim 2 \text{ Tbit/s}$) switching network equipment.

This approach was kept for DAQ2, with some minor modifications: the network between the front end readout link and the readout units is not foreseen to be a full switching matrix any more (i.e. in the past any readout link device could be routed to any readout unit PC) but merely an aggregation layer (each readout link device can only send to a small number of readout unit

PCs). Dimensioning the readout unit capacities larger than the design bandwidth allows for some level of fault tolerance in case e.g. a readout unit PC or a laser transmitter stops working. The possibility of adding a second layer of Ethernet switches between the FEROLs and the RUs to build a full mesh network which will further improve fault tolerance during data taking is under study.

5. Performance considerations

This section describes some performance issues we have encountered while testing high speed links and highlights some advantages of Infiniband over TCP/IP.

5.1. TCP/IP

Running TCP/IP communication at a speed of 40 Gbit/s (i.e. more than a factor 10 higher than DAQ1) requires some level of performance tuning. The first set of system parameters to adjust is to increase the Linux kernel TCP socket buffer settings. The default values are too small for 40 Gbit/s operation. Larger socket buffers on the receive side will allow the receiver to send fewer acknowledgments and thus put less load on the CPU.

A further important option is to enable the support for ‘large receive offload’ which most modern network cards support. When this is enabled, the network card aggregates multiple received packets into a single large buffer before an interrupt is generated. This further reduces the load on the CPU.

High speed network cards have multiple receive queues each of which corresponds to a separate interrupt number. Each interrupt can be sent to a specified set of CPU cores. Even though the operating system has a mechanism to automatically balance the interrupts across CPU cores (through the `irqbalance` daemon), we found that for our application assigning the interrupts to cores ‘by hand’ can lead to better performance and gives better control over the system.

5.2. Infiniband

Infiniband [6] was designed as a high speed interconnect for data centers. It has a transport layer which can be implemented in hardware. Contrary to the BSD socket system calls, data buffers passed to the sending functions may only be released after acknowledgment by the remote end. This eliminates the need for copying the data from user space and into a kernel buffer as is done with the BSD sockets interface.

These features allow to reduce the CPU load on the participating hosts. For simple applications however, transferring data over Infiniband requires a larger number of library calls to setup the data transfer than the sockets API.

The fastest links currently available commercially are FDR (14 Gbit/s) links with four parallel lanes corresponding to a line speed of 56 Gbit/s. According to the TOP500 web site [7] which ranks the 500 most powerful computer systems in the world, Infiniband had a share of 41% of the ranked systems in June 2013 while the largest systems share (27%) in November 2002 (shortly before the release of the DAQ1 technical design report) was held by Myrinet (see figure 2).

5.3. Memory bandwidth

With the introduction of the Nehalem microarchitecture, Intel CPUs have a per CPU memory controller rather than an external Memory Controller Hub (MCH) which was shared between CPUs in previous architectures. This implies that each CPU can access only part of the memory directly through its memory controller while the remaining part must be accessed via the Quick Path Interconnect (QPI) to the other CPU in a dual processor system. If optimal performance is required, care has to be taken to allocate memory on the correct CPU.

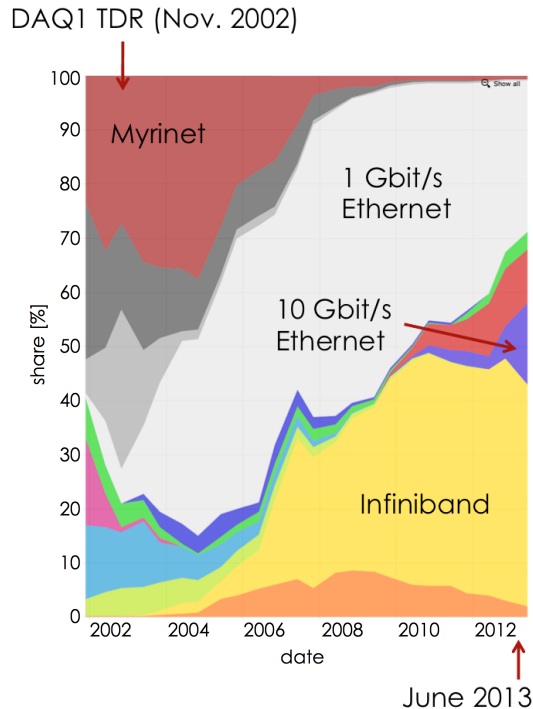


Figure 2. Fraction of supercomputer systems built using a given interconnect family in the Top 500 supercomputer list as function of time [7]. The region at the left end corresponds approximately to the time of the publication of the DAQ1 Technical Design Report (November 2002) while the region at the right end corresponds to the situation in June 2013.

6. DAQ2 test setup and results

In order to develop software and test the hardware and software performance, we have set up a test bed which allows to test various configurations and possible variants of the DAQ2 setup (see figure 3).

We have evaluated the performance of a 15 readout unit by 15 builder unit system interconnected by an Infiniband network and have seen that we meet our performance requirements (see figure 4). More importantly, we tested the evolution of the performance for 1 by 1, 2 by 2 up to 15 by 15 and the result makes us confident that the full scale system will have similar per node throughput than the test system.

To verify the performance of a slice of the chain from FEROLs up to the builder units, we have measured the throughput as function of the fragment size from 12 FEROLs to one readout unit to four builder units (figure 5). The readout rate crosses the 100 kHz line typically at a fragment size of 4 kByte in this configuration (aggregation of 12 streams). Note that the current design of DAQ2 is such that most readout units will aggregate less than 12 streams.

7. Conclusions

CMS has designed a central data acquisition system for post-long shutdown 1 data taking replacing outdated standards by modern technology. This system has about twice the event building capacity of the DAQ system for LHC run 1. It can accommodate a large dynamic range of up to 8 kByte fragments per front end driver and is flexible enough to handle small fragments from other front end drivers at the same time.

The increase in networking bandwidth over recent years was faster compared to the increase in event sizes we expect for LHC run 2. This allows us to reduce the number of event builder PCs by an order of magnitude while each PC must handle a factor of about ten more bandwidth. With a small scale demonstrator, we have performed various performance tests and demonstrated that we master the selected high-speed-networking technologies. The results from scaling tests make us confident that the performance requirement will be met with the full scale DAQ2 system.

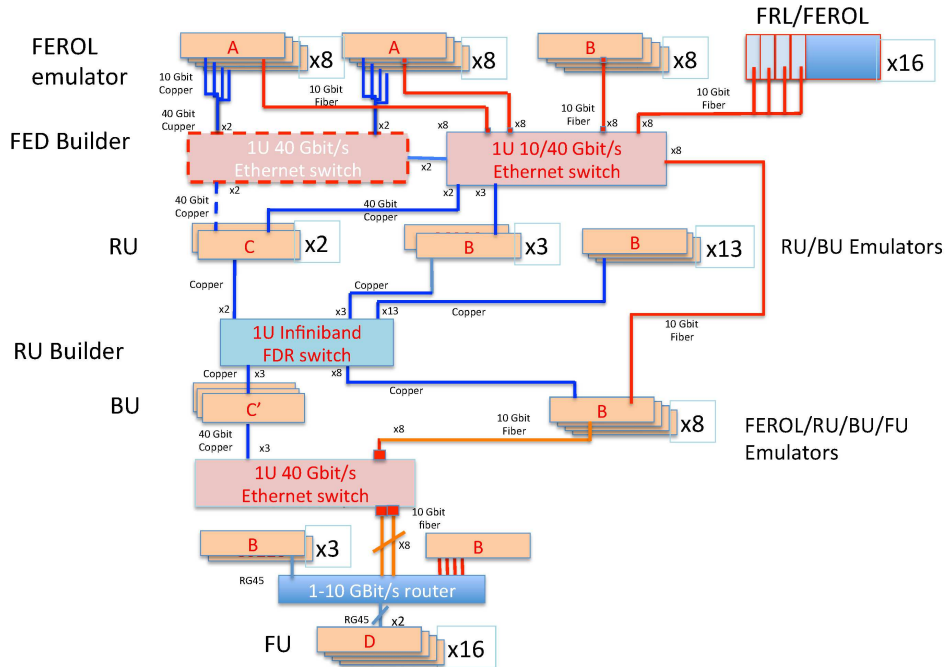


Figure 3. Test setup for development and performance testing for DAQ2. The boxes labeled A through D correspond to different types of PCs, the big labels next to the boxes indicate their function. The larger boxes represent network switches and the lines show the network connections. Event data flows from top to bottom in general.

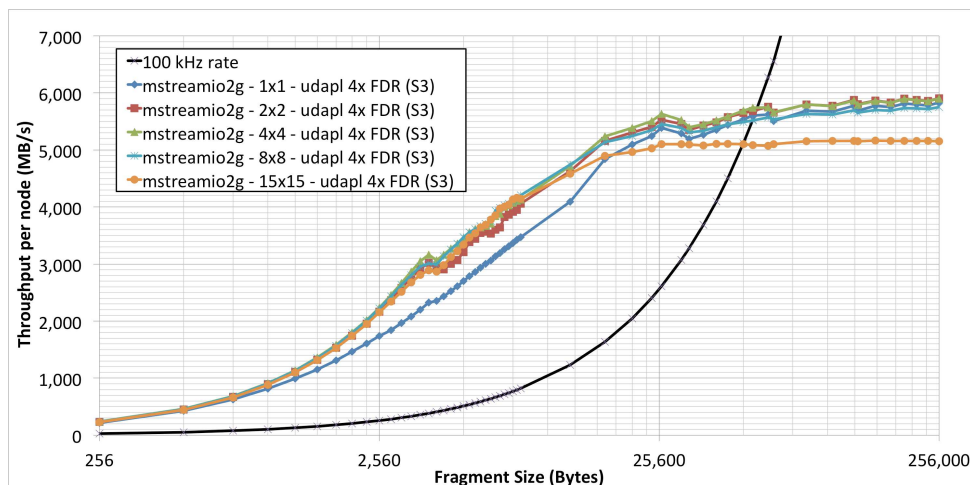


Figure 4. Throughput across an Infiniband network for N PCs sending data to N receiving PCs as function of the fragment size, using the `mstreamio2g` benchmark and the `uDAPL` library [8]. The exponentially growing (black) line corresponds to a 100 kHz readout rate. The bulleted (colored) lines correspond to throughput measurements with $N = 1, 2, 4, 8, 15$ respectively.

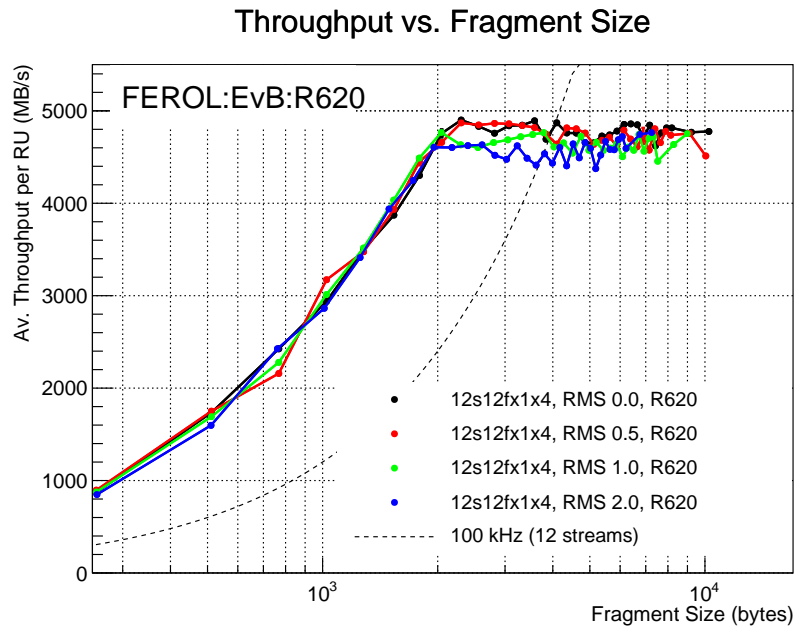


Figure 5. Throughput for a single readout unit receiving data from 12 FEROLs (sending one TCP stream each) and sending it to 4 Builder Unit PCs. The fragment sizes are drawn from a log-normal distribution. The value on the horizontal axis corresponds to the mean of the log-normal distribution, the different lines to different widths of the distribution. The dashed line indicates the throughput corresponding to a 100 kHz readout rate at the given average fragment size. The 100 kHz rate can be sustained up to about 4 kByte average fragment size.

The installation activities for DAQ2 have started already, the full deployment is foreseen for mid 2014.

8. Outlook

For the High Luminosity LHC (expected to start in 2024 after the long shutdown 3), the CMS central DAQ will likely have to build 4 MBytes large events at a readout rate of 1 MHz, corresponding to a processing capacity of 32 Tbit/s. While this seems still challenging and expensive today, we expect this to be feasible due to the reduction of the cost/performance ratio of computing equipment and further increases in networking link speeds in the next decade.

Acknowledgments

This work was supported in part by the DOE and NSF (USA) and the Marie Curie Program.

References

- [1] Bauer G, Behrens U, Bowen M, Branson J, Bukowiec S *et al.* 2012 *J.Phys.Conf.Ser.* **396** 012007
- [2] Micro Telecommunications Computing Architecture base specification: Micro TCA <http://www.picmg.org/>
- [3] Racz A, McLaren R and van der Bij E <http://hsi.web.cern.ch/HSI/s-link>
- [4] Zejdl P (these proceedings)
- [5] Mommsen R K (these proceedings)
- [6] InfiniBand Architecture Specification Release 1.2.1 <http://www.infinibandta.org/>
- [7] Top500 <http://www.top500.org/>
- [8] User Direct Access Programming Library, <http://www.datcollaborative.org/udapl.html>